World Journal on Educational Technology: Current Issues

Volume 14, Issue 5, (2022) 1518-1528

www.wj-et.eu

Teaching students to use Decision trees (Dt) for unstructured data

Konstantin Bogdanov *, Department of Informatics and Computer Engineering, Reshetnev Siberian State University of Science and Technology, Krasnoyarsky Rabochy Av., 31, Krasnoyarsk, 660037, Russian Federation, https://orcid.org/0000-0002-9189-0517

Dmitry Gura, Department of Cadastre and Geoengineering, Kuban State Technological University, Moskovskaya str., 2B, Krasnodar, 350072, Russian Federation; Department of Geodesy, Kuban State Agrarian University, Kalinina Str., 13, Krasnodar, 350044, Russian Federation https://orcid.org/0000-0002-2748-9622

Dustnazar Khimmataliev, Department of Pedagogy, Chirchik State Pedagogical Universitete, Oyimarik str., 7/2, Chirchik, 100124, Uzbekistan, https://orcid.org/0000-0002-0663-8473

Yulia Bogdanova, Department of Foreign Languages, Federal State Budgetary Educational Institution of Higher Education Northern Trans-Ural State Agricultural University, Respubliki str., 7, Tyumen, 625003, Russian Federation, https://orcid.org/0000-0002-7256-1590

Suggested Citation:

Bogdanov, K, Gura, D., Khimmataliev, D. & Bogdanova, Y. (2022). Teaching students to use Decision trees (Dt) for unstructured data. *World Journal on Educational Technology: Current Issues*. 14(5), 1516-1528. https://doi.org/10.18844/wjet.v14i5.7335

Received from February 15, 2022; revised from July 10, 2022; accepted from September 24, 2022.
Selection and peer review under responsibility of Prof. Dr. Servet Bayram, Medipol University, Turkey

Abstract

The research aims to analyze the importance of teaching to use unstructured data methods that students generate from the learning activities and examine the relative efficiency of the decision trees within load conditions and self-efficacy of each learner. The present research collected the data using a questionnaire to analyze self-efficacy and cognitive load among students. The sample included 150 students divided into two groups. The research revealed no significant differences in self-efficacy between the two groups participants ($F = 0.01$, $p > 0.05$). According to the results, no differences were identified between the students who worked with unstructured data using decision trees and those students who analyzed the unstructured data using association rules. The research uses an independent t-test for the analysis of cognitive load within the academic environment. No significant differences were detected concerning cognitive load between the two groups of participants.

Keywords: unstructured data, decision trees, association rules, self-efficacy, cognitive load, SDGs.

*ADDRESS OF CORRESPONDENCE: Dmitry Gura, Kuban State Technological University Moskovskaya Ulitsa, 2, Krasnodar, Krasnodarskiy kray, Russia, 350072
Email address: dmit_gura@rambler.ru

Introduction

Unstructured data means any information that does not have a predefined dataset. It usually includes texts with dates, numbers, and facts stored in an unstructured format. The data generated in textual or non-textual format is difficult to analyze, in particular, using standard programs designed for structured data analysis (Gandomi & Haider, 2015).

In 1998, Merrill Lynch had claimed that approximately 81-90% of the important business data had been in an unstructured format (Sims et al., 2017). Nevertheless, no statistical or quantitative research was conducted at that time to examine the issue (Fang et al., 2015). Years later, Computerworld estimated the amount of unstructured data in global organizations and found that 70-80% of the world's data was unstructured (Khan & Vorley, 2017).

Unstructured data management solutions have remained one of the main challenges in the information technology (IT) industry (Lněnička et al., 2021). The analytical tools and methods used to perform structured data analysis are not efficient for unstructured data. New approaches are required to analyze large unstructured datasets (Eberendu, 2016). A decision trees analytical tool becomes one of the popular methods used for automated data analysis including the example-based learning approach (Al-Barrak & Al-Razgan, 2016; Segatori et al., 2017).

The decision tree models are based on a multi-level structure including elements such as (Menezes et al., 2022): root nodes and branches (the topmost nodes) - attributes expressed in a descriptive way; leaf nodes (the lowest nodes) - classes that characterize a certain subject; connections between nodes and leaves. This method is based on a hierarchical structure, where nodes represent entities and objects, and solutions. The classification starts at the root node, achieves the leaves, checks the values of the topmost nodes. Nevertheless, in some cases, it is not possible to classify an object based on its properties only (Dziwiński et al., 2018). Fuzzy logic is used to analyze complex problems approached from multiple and competing, perspectives. The object belonging to a certain class is important in the unstructured data analysis. Following the fuzzy decision trees, an object can have properties of several attributes. For each attribute, it is necessary to define several linguistic values and the object belonging. A fuzzy decision tree treats data according to its belonging to certain classes instead of analyzing them as a set of objects of a single node (Dziwiński et al., 2018).

Today, large datasets are stored in an unstructured database format. The decision trees method is one of the most popular in teaching students to analyze textual and non-textual information. The research is needed to examine the impact of teaching using decision trees methods on academic achievements and identify the negative and positive outcomes of learning (Izza et al., 2020; Kamiński et al., 2018). It will help to improve the learning process and develop a favorable learning environment for students.

Literature review

With the arrival of big data, the volume of unstructured data has increased significantly. Decision trees methods are one of the most widely used classification models to structure large datasets. Unstructured data including texts should be converted to a structured format for decision trees analysis (Yang, 2019). The scholars developed a framework to apply decision trees for datasets with unstructured data (King, 2015). The CUST decision tree model was proposed to analyze unstructured data. CUST is based on the partitioning criteria, generated by unstructured attribute values. It reduces significantly the number of datasets scans by using proper data structures. Tests confirmed that CUST improves the efficiency of classification schemes for unstructured data (Mittal et al., 2017).

The era of big data optimizes the business. Machine learning has become a popular method for processing data. Data can be stored in different formats including textual formats, images, audio,

video, and signals. Machine learning is aimed to investigate and predict output values based on input data (Liu et al., 2016). In big datasets, there is a large number of functions that can lead to a large number of irrelevant functions. Most machine learning algorithms are sensitive to irrelevant functions, so evaluation and selection of functions are very important for machine learning. The functions assessment can help to identify the datasets to be extracted from unstructured data (Savage & Yuan, 2016).

Other scholars focused on the image processing techniques in classification (Liu et al., 2017). The researchers paid special attention to the two types of function selection, namely the filter and wrapper methods. The research examined several approaches to machine learning that were widely used in image classification and helped to identify the limitations of the proposed algorithms for functions evaluation. The experimental research examined the performance of C 4.5 (decision tree learning algorithm) and other algorithms (Naive Bayes, K Nearest Neighbors, and Multi-Layer Perceptron) on five image datasets of the UCI repository (Wang & Yu, 2017). The decision tree learning models have been examined to analyze the learning models in terms of performance evaluation in the training phase. The analysis helped to understand how rules derived from the decision tree could be used to evaluate functions in the validation phase (Aghabozorgi et al., 2015). The impact of these methods on teaching and the effectiveness of learning has not been investigated yet.

Using fuzzy decision trees, Russian scholars have developed a model for electronic unstructured text documents, taking into account syntactic relationships and functions of words in sentences (Dli et al., 2019). A large number of electronic text messages (complaints, appeals, suggestions, etc.) in an unstructured format are posted on the Internet portals of the state agencies. The quality and speed of automated processing of such requests depend greatly on assigning text to certain classes (subject area) (Dli et al., 2018). Distinctive features of these messages including small size, errors, inadequate structure prevent proper text documents categorization.

The construction of a decision tree is based on the analysis of vocabularies to determine the certain class to be associated with the data points and the distances between classes in n-dimensional feature space (Chen et al., 2017). The scholars underlined that this model helped to classify unstructured electronic text documents with interrelated headings and improved document processing. The research did not investigate possible ways to use the proposed model for other types of unstructured data.

Efficient training of decision trees has been examined (Vos & Verwer, 2021). The existing approaches for decision tree learning are expensive and require time to be completed. The scholars proposed the GROOT algorithm, an efficient algorithm for training robust decision trees that runs in seconds or minutes. The findings of single trees and ensembles on 14 structured datasets, as well as on MNIST and Fashion-MNIST analysis demonstrated that GROOT performed faster than other analytical models and demonstrated the best results for computation accuracy on structured data (Andriushchenko & Hein, 2019). The researchers did not examine the possible disadvantages of using this algorithm.

A new hybrid method for online learning has been explored by the authors. It combines classification models such as incremental decision trees (ITI-2.8), fuzzy logic for conceptual learning and appropriate reasoning to accept noisy and imprecise input data (Iswanto et al., 2016). The algorithm proposed in this article is based on three new aspects. The algorithm introduced a fuzzy associative memory (FAM) system, defined as clusters of grouped fuzzy decision trees (FDT). FAMs have generated automatically, interactively and incrementally. In real-time, automated unstructured data collection is possible to perform (Batra & Agrawal, 2018). Each FAM, including a system of Multiple Inputs and Single Output (MISO), develops a unique model developing online and based on the rewarding actions that the robot experiences. Fuzzy data vectors are inserted online in ITI-2.8 as

they are collected incrementally for knowledge extraction. The FAM development and growth since its inception is fully automated and does not require expert support. Based on the decision trees root nodes, the FAM can grow rapidly keeping the clarity and expressiveness of the fuzzy rules generated by the decision tree. The fuzzy logic allows to blend of different behaviour patterns. The global path planning is constructed by switching between the local FAMs (Ahmed et al., 2020). The main research limitation was the novelty of the proposed algorithm and a lack of data on its implementation cases. Further research is needed to investigate the application of the algorithm in different contexts.

The researchers examined the use of decision trees, classification and regression trees (CART) and boosted regression trees (BRT) to understand missing values in the data structure (Tierney et al., 2015). For the analysis, the data was collected from workers at three different industrial enterprises in Australia. There were 7915 observations included. The proposed approach was analyzed using an occupational health dataset including questionnaires, medical tests and environmental monitoring. Simulation research investigated decision tree models for data analysis of the versions with artificially inserted missing values (James et al., 2021; Pallant, 2020). The CART and BRT models were effective for analysis of the missing data in specific data types (healthcare or environmental), missing data on the data collection place, the number of visits, and factors affected by extreme values. The simulation has shown that CART models can be used to identify the variables and values being responsible for missing data. The missing value in each variable for unstructured data was greater than for structured ones. It was found that the CART and BRT models can be used for descriptive purposes in missing datasets (Nguyen et al., 2017). CART models are more efficient than BRT models for missing data methods for exploratory data analysis and selection of values important for predicting missing values. BRT models show that missing values present in the dataset can impact the performance of the model and the consequences of missing data for bias in estimates of causal effects depending on the type of variable that is missing (Pampaka et al., 2016). The scholars recommend the CART and BRT models for analysis and understanding the missing data. The online decision trees were used to support the self-efficacy of students in the research laboratory (McLean et al., 2020). Researchers often report experimental failure; however, many laboratories rely on established protocols to ensure students can get trustworthy results.

The research laboratory does not provide students with all the necessary tools to conduct their experiments. Nevertheless, it provides students with the opportunity to experience *the research failure* in a safe environment as a part of the problem-solving skills development process (Bartimote-Aufflick et al., 2016). Academic institutions ensure a safe space where students are not discouraged by failure and resilience. Online decision trees have been designed to help students to use the laboratory protocols and give them feedback. Thus, students are encouraged to develop the laboratory protocol for their research (Sebastián et al., 2021). Online decision trees can be described as a scenario followed by students. The students choose options and go through different paths to achieve different results in their experiments. They receive feedback and tutorials throughout the simulation directed by choice options (Cooper et al., 2018).

The new approach helps students to develop problem-solving skills and gain theoretical knowledge on different research phases. The current research aims to assess how online decision trees affect student self-efficacy, metacognition, and motivation to conduct laboratory experiments. The proposed approach was based on blended methods. During the academic semester, three surveys were conducted. The findings revealed that online decision trees introduced before the laboratory work improved students' self-efficacy and intrinsic motivation. Nevertheless, extrinsic motivation and metacognition remained unchanged (Dohn et al., 2016).

Setting goals

The research aims to improve teaching unstructured data to students and develop a good learning environment for them. *The purpose* is to examine the importance of teaching unstructured data methods to students and examine the relative efficiency of the decision trees within load conditions and the self-efficacy of each learner. *The objectives* included the following:

- determine the level of self-efficacy and cognitive load in students;
- identify significant differences in self-efficacy and cognitive load in the control and experimental groups.

Methods and materials

Quasi-experimental research was conducted in *the Information Technology* course to analyze the role of teaching unstructured data methods to students and examine the relative efficiency of the decision trees within load conditions and the self-efficacy of each learner. The quasi-experimental design provides a scientific approach to research because it examines the causal relationships between the independent and dependent variables within a controlled environment. For some experimental research, quasi-experimental research is the best approach to develop control methods and minimize risk factors affecting the research validity.

Study participants

The research involved 2nd-year students of the [BLINDED] University [BLINDED] and the [BLINDED] University. The sample consisted of 150 students: the experimental group included 75 randomly selected students (35 women and 40 men) and the control group included 75 students (40 women and 35 men). The average age is 19 years. The experimental group used decision trees to analyze unstructured data and the control group used association rules to examine unstructured data.

Research tools

The research used a questionnaire approach to analyze performance measures of self-efficacy and cognitive load in participants (Appendix 1-2). The Likert scale was used in the questionnaire. The scale contained 5 responses options with two extreme sides and a neutral opinion. Instead of “strongly agree” or “strongly disagree,” the assessment is based on a numerical description, using 1 to 5 points to evaluate the answers. The self-efficacy questionnaire has been revised (George & Mallery, 2003). It consisted of eight items based on a 5-point Likert scale. The cognitive load questionnaire was also revised. It consisted of eight items using a 7-point Likert scale and was divided into two parts.

Statistical analysis

Cronbach's alpha was used to test the research reliability. According to George and Mallery (2003), the Cronbach's alpha scale assigns the following values: > 0.9 - excellent; > 0.8 - good; 0.7 - acceptable; 0.6 - doubtful; and > 0.5 - bad. The Cronbach’s alpha value for the self-efficacy was 0.92, and 0.96 for cognitive load in students. The questionnaire's reliability was high, and a survey was conducted. The statistical data was assessed using an analysis of covariance (ANCOVA).

Moreover, the Shapiro-Wilk test was used to assess the data. The result of this test was 0.98 ($p > 0.05$). It indicated a normal distribution of the data. The Leuven test was performed to examine the uniform distribution variance ($F = 1.65$, $p > 0.05$). It highlights that the assumption is reasonable and no significant differences are found between the two groups. Homogeneity of regression slopes was also confirmed, with the possibility to perform the one-way ANCOVA ($F = 0.26$, $p > 0.05$). The cognitive load of students was analyzed using a t-test.

Results

Self-efficacy analysis

Table 1 includes the results of ANCOVA analysis on self-efficacy in students. The adjusted mean and standard error were 3.28 and 0.12 for the experimental group and 3.27 and 0.11 for the control group, respectively.

Table 1 Analysis of self-efficacy in students (ANCOVA)

Group	#	Mean	SD	Average mean	SE	F
The experimental group	75	3.29	0.92	3.38	0.12	0.01
The control group	75	3.27	0.69	3.27	0.11	

No significant differences between the two groups were identified (F = 0.01, p> 0.05). It means that there is no significant difference in the self-efficacy in students learning the unstructured data with decision trees and those who used the association rules. Students in general had a low level of self-efficacy, which is evidence of psychological unpreparedness to master the curriculum. The formation of students' self-efficacy can be facilitated by such conditions under which they will have the opportunity to gain valuable experience in the development and evaluation of professional competencies in interaction with all educational process subjects. An important condition for successful professionalization is the inclusion of self-efficacy components, in particular professional abilities, into the value-motivational schemes of students' professional and personal self-development. The forms and methods of professional training activity, such as communicative training, mutual evaluation of each other's academic achievements during seminars based on microgroup interaction, organization of dyads and triads during practical and laboratory classes also contribute to professional self-development.

Cognitive load analysis

A t-test was used to analyze the cognitive load in students. Table 2 includes the results of the t-test assessing the cognitive load.

Table 2 Cognitive load analysis using independent t-criterion

Cognitive load	**Group**	**N**	**Mean**	**SD**	**m**
Cognitive load	The experimental group	75	3.02	1.56	0.70
	The control group	75	2.97	1.34	
Cognitive efforts	The experimental group	75	3.48	1.52	-0.55
	The control group	75	3.40	1.53	
Total	The experimental group	75	3.34	1.55	0.74
	The control group	75	2.28	1.25	

The mean and standard deviations of the assessments were 3.34 and 1.55 for the experimental group students and 3.28 and 1.25 for the control group students. No significant differences were found between the two groups in cognitive load. Understanding that other factors may influence the results of the experiments conducted (students' prior knowledge, experience with computers, teaching quality, gender differences, etc.), the authors deliberately limited the scope of the study to show the need to predict students' cognitive load when using electronic resources. The use of the developed teaching methods makes it possible to control the cognitive load, in particular focusing on the distribution of students in the group according to their learning advantages (Chu et al., 2015). Thus, the established correlations between teaching methods and the cognitive load they

experience when working with electronic resources will be useful for analyzing the effectiveness and refinement of teaching methods. By consciously combining informational educational resources developed with consideration for psychological and pedagogical features of knowledge perception, a teacher gets an opportunity to optimize students' learning activities and improve the quality of learning.

Discussion

The research found no significant differences in self-efficacy and cognitive load between the experimental and control groups participants. The researchers explain low self-efficacy in students by a lack of teaching time spent on decision trees. It was found that short teaching time influenced student self-efficacy. Students need more time to develop their cognitive skills. The experimental group students used this method for the first time, so a lack of learning experience was also an important factor of low self-efficacy results. The scholars admit that the introduction of new technologies is still largely unrealized, therefore, students should spend more time on self-efficacy. New research respective are identified and future research is needed to investigate students' self-efficacy in different contexts. The research on self-efficiency is not accurate enough because it has some limitations including the experimental method, measuring methods, and time.

Students using decision trees and association rules experienced the same cognitive load. The scholars admit that future research should overcome the limitations of the experimental method, the measuring method, and time constraints. The research does not provide insights on the impact of teaching students decision trees in unstructured data analysis and relations with self-efficacy and cognitive load. The comparison of the results is problematic.

The research (McLean et al., 2020) analyzed the use of online decision trees to support student self-efficacy in the research. The study assessed the effectiveness of online decision trees in metacognition, motivation, and self-efficacy. The research revealed that students reported higher levels of self-efficacy and intrinsic motivation working on the lab assignments. Low self-esteem and self-efficacy were reported by students using the laboratory protocol. The research concluded that interactive learning was effective and should be widely used in unstructured data management (Williams & Rhodes, 2016).

The research investigates the application of decision trees in teaching university students. The research identified ways to improve self-esteem in students in higher education. Using the decision tree, the scholars developed a framework to improve students' self-esteem. These decision rules help to intensify education. The success of education and the development of professional competencies largely depend on the level of self-esteem. Self-esteem has a marked effect on academic performance and increases the desire to learn. The researchers claim that educators should establish and maintain healthy environments for students to learn and grow. Educational institutions can also play an important role in referring students experiencing low self-esteem to professionals.

Conclusions

At present, data are stored in an unstructured format. Decision trees are one of the most widely used methods in teaching students to work with different types of information including unstructured data. The research highlights that there is a need to investigate the role and impact of using decision trees in unstructured data analysis. The research examines the relative efficiency of the decision trees within load conditions and the self-efficacy of each learner. The results revealed the level of self-efficacy and cognitive load of students, as well as differences in self-efficacy and cognitive load in the control and experimental groups. The research found no significant differences between the two groups of participants ($F = 0.01$, $p > 0.05$). The research suggested that there was no significant difference in self-efficacy in students working with decision trees and association rules

applied to unstructured datasets. An independent t-test found no significant differences between the two groups in students' cognitive load.

The research limitations include a lack of analysis of large samples. Future research is needed for long time perspectives to assess the teaching of the unstructured data to students. There is a need to examine the sustainability of the research results. Moreover, the sample should be increased to assess much more students and ensure the accuracy of the results. Finally, factors such as different learning styles, different characteristics, academic performance, and gender can also be considered to expand the scope of the research.

The findings can be used by educators to develop a framework of teaching students to work effectively with unstructured data. No significant differences were identified concerning the impact of both methods on self-efficacy and cognitive load in students. Both approaches can be successfully used in teaching. Future research is required to explore the positive and negative aspects of using decision trees in learning for classification and regression tasks.

References

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information systems*, *53*, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Ahmed, S., Shekha, M., Skran, S., & Bassyouny, A. (2022). Investigation of optimization techniques on the elevator dispatching problem. *arXiv preprint arXiv:2202.13092*. https://doi.org/10.48550/arXiv.2202.13092

Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: A case study. *International Journal of Information and Education Technology, 6*(7), 528–533. https://doi.org/10.7763/IJIET.2016.V6.745

Andriushchenko, M., & Hein, M. (2019). Provably robust boosted decision stumps and trees against adversarial attacks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (13017–13028). ACM. Retrieved from https://dl.acm.org/doi/10.5555/3454287.3455453

Bartimote-Aufflick, K., Bridgeman, A., Walker, R., Sharma, M., & Smith, L. (2016). The study, evaluation, and improvement of university student self-efficacy. *Studies in Higher Education*, *41*(11), 1918–1942. https://doi.org/10.1080/03075079.2014.999319

Batra, M., & Agrawal, R. (2018). Comparative analysis of decision tree algorithms. In *Nature inspired computing* (pp. 31–36). Springer. https://doi.org/10.1007/978-981-10-6747-1_4

Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, *72*, 221–230. https://doi.org/10.1016/j.eswa.2016.10.065

Chu, H. C., Yang, K. H., & Chen, J. H. (2015). A time sequence-oriented concept map approach to developing educational computer games for history courses. *Interactive Learning Environments*, *23*(2), 212–229. https://doi.org/10.1080/10494820.2014.979208

Cooper, K. M., Krieg, A., & Brownell, S. E. (2018). Who perceives they are smarter? Exploring the influence of student characteristics on student academic self-concept in physiology. *Advances in Physiology Education, 42*(2), 200–208. https://doi.org/10.1152/advan.00085.2017

Dli, M., Bulygina, O. V., & Kozlov, P. Y. (2018). Development of multimethod approach to rubrication of unstructed electronic text documents in various conditions. In *2018 International Russian Automation Conference (RusAutoCon)* (pp. 1–5). IEEE. https://doi.org/10.1109/RUSAUTOCON.2018.8501815

Dli, M., Bulygina, O., & Kozlov, P. (2019). Application of fuzzy decision trees for rubricating unstructured electronic text documents. In *Proceedings of the IS-2019 Conference* (pp. 108–118). CEUR. Retrieved from http://ceur-ws.org/Vol-2475/paper9.pdf

Dohn, N. B., Fago, A., Overgaard, J., Madsen, P. T., & Malte, H. (2016). Students' motivation toward laboratory work in physiology teaching. *Advances in Physiology Education, 40*(30), 313–318. https://doi.org/10.1152/advan.00029.2016

Dziwiński, P., Bartczuk, Ł., & Przybyszewski, K. (2018, June). A population based algorithm and fuzzy decision trees for nonlinear modeling. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 516–531). Springer. https://doi.org/10.1007/978-3-319-91262-2_46

Eberendu, A. C. (2016). Unstructured Data: An overview of the data of Big Data. *International Journal of Computer Trends and Technology*, *38*(1), 46–50. https://doi.org/10.14445/22312803/IJCTT-V38P109

Fang, Y., Hoang, T. T., Becchi, M., & Chien, A. A. (2015, December). Fast support for unstructured data processing: The unified automata processor. In *Proceedings of the 48th International Symposium on Microarchitecture* (pp. 533–545). ACM. https://doi.org/10.1145/2830772.2830809

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

George, D., & Mallery, P. (2003). *SPSS for windows*: *Step by step*. Allin & Bacon. https://dl.acm.org/doi/abs/10.5555/557542

Iswanto, I., Wahyunggoro, O., & Cahyadi, A. I. (2016). Path planning based on fuzzy decision trees and potential field. *International Journal of Electrical and Computer Engineering*, *6*(1), 212–222. https://doi.org/10.11591/ijece.v6i1.8606

Izza, Y., Ignatiev, A., & Marques-Silva, J. (2020). On explaining decision trees. *arXiv preprint arXiv:2010.*11034. https://doi.org/10.48550/arXiv.2010.11034

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Statistical learning. In *An introduction to statistical learning* (pp. 15–57). Springer. https://doi.org/10.1007/978-1-0716-1418-1_2

Kamiński, B., Jakubczyk, M., & Szufel, P. (2018). A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research, 26*, 135–159. https://doi.org/10.1007/s10100-017-0479-6

Khan, Z., & Vorley, T. (2017). Big data text analytics: An enabler of knowledge management. *Journal of Knowledge Management*, *21*(1), 18–34. https://doi.org/10.1108/JKM-06-2015-0238

King, M. A. (2015). *Ensemble learning techniques for structured and unstructured data.* Doctoral dissertation. Virginia Polytechnic Institute and State University. Retrieved from https://www.proquest.com/openview/0a597b0692d3a0a8b4939ce869cccba2/1?pq-origsite=gscholar&cbl=18750

Liu, H., Cocea, M., & Ding, W. (2017). Decision tree learning based feature evaluation and selection for image classification. In *2017 International Conference on Machine Learning and Cybernetics (ICMLC)* (pp. 569–574). IEEE. https://doi.org/10.1109/ICMLC.2017.8108975

Liu, H., Gegov, A., & Cocea, M. (2016). *Rule based systems for big data: A machine learning approach*. Springer. https://doi.org/10.1007/978-3-319-23696-4

Lněnička, M., Machova, R., Volejníková, J., Linhartová, V., Knezackova, R., & Hub, M. (2021). Enhancing transparency through open government data: The case of data portals and their

features and capabilities. *Online Information Review, 45*(6), 1021–1038. https://doi.org/10.1108/OIR-05-2020-0204

McLean, S., Meadows, K. N., Heffernan, A., & Campbell, N. (2020). Using online decision trees to support students' self-efficacy in the laboratory. *Advances in Physiology Education*, *44*(3), 430–435. https://doi.org/10.1152/advan.00016.2019

Menezes, A. G., Araujo, M. M., Almeida, O. M., Barbosa, F. R., & Braga, A. P. (2022). Induction of decision trees to diagnose incipient faults in power transformers. *IEEE Transactions on Dielectrics and Electrical Insulation*, *29*(1), 279–286. https://doi.org/10.1109/TDEI.2022.3148453

Mittal, K., Aggarwal, G., & Mahajan, P. (2017). A comparative study of association rule mining techniques and predictive mining approaches for association classification. *International Journal of Advanced Research in Computer Science*, *8*(9), 365–372. https://doi.org/10.26483/ijarcs.v8i9.4984

Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model checking in multiple imputation: An overview and case study. *Emerging Themes in Epidemiology*, *14*(1), 1–12. https://doi.org/10.1186/s12982-017-0062-6

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. Routledge. https://doi.org/10.4324/9781003117452

Pampaka, M., Hutcheson, G., & Williams, J. (2016). Handling missing data: Analysis of a challenging data set using multiple imputation. *International Journal of Research & Method in Education*, *39*(1), 19–37. https://doi.org/10.1080/1743727X.2014.979146

Savage, R. S., & Yuan, Y. (2016). Predicting chemoinsensitivity in breast cancer with omics/digital pathology data fusion. *Royal Society Open Science, 3*(2), 140501. https://doi.org/10.1098/rsos.140501

Sebastián, G., Tesoriero, R., & Gallud, J. A. (2021). Unified abstract mechanism to model language learning activities. *Computing and Informatics, 40*(2), 249–276. https://doi.org/10.31577/cai_2021_2_249

Segatori, A., Marcelloni, F., & Pedrycz, W. (2017). On distributed fuzzy decision trees for big data. *IEEE Transactions on Fuzzy Systems*, *26*(1), 174–192. https://doi.org/10.1109/TFUZZ.2016.2646746

Sims, R. R., Erickson, E. H., & Erickson, J. P. (2017). European enterprise information portals and global communications. In *The new millennium: Challenges and strategies for a globalizing world* (pp. 225–239). Routledge. Retrieved from https://www.taylorfrancis.com/chapters/edit/10.4324/9781315187181-12/european-enterprise-information-portals-global-communications-ronald-sims-eric-erickson-jeffrey-erickson

Tierney, N. J., Harden, F. A., Harden, M. J., & Mengersen, K. L. (2015). Using decision trees to understand structure in missing data. *BMJ Open, 5*, e007450. http://dx.doi.org/10.1136/bmjopen-2014-007450

Vos, D., & Verwer, S. (2021). Efficient training of robust decision trees against adversarial examples. In *International Conference on Machine Learning* (pp. 10586–10595). PMLR. Retrieved from http://proceedings.mlr.press/v139/vos21a/vos21a.pdf

Wang, Y., & Yu, H. (2017). Facial expression-aware face frontalization. In *LNCS Proceedings of Asian Conference on Computer Vision* (pp. 375–388). Springer International Publishing. Retrieved

from https://www.springerprofessional.de/en/facial-expression-aware-face-frontalization/12134034

Williams, D. M., & Rhodes, R. E. (2016). The confounded self-efficacy construct: Conceptual analysis and recommendations for future research. *Health Psychology Review*, *10*(2), 113–128. https://doi.org/10.1080/17437199.2014.941998

Yang, F. J. (2019, December). An extended idea about decision trees. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 349–354). IEEE. https://doi.org/10.1109/CSCI49370.2019.00068

Appendix 1

Self-efficacy assessment questionnaire

Please rate the degree to which you agree or disagree with each of the following, from 1 to 5 (where 1 - *strongly agree*, 5 - *strongly disagree*):

(1) I will get an excellent mark after training.

(2) I can understand the most difficult material presented in teaching unstructured data.

(3) I can understand the basic concepts of unstructured data.

(4) I can understand the most difficult material.

(5) I can do a great job with the unstructured data assignments and tests.

(6) I expect to perform well using unstructured data.

(7) I can master the required skills.

(8) I will do my best and master the skills in spite of some difficulties that may arise working with unstructured data.

Appendix 2

Cognitive Load Assessment Questionnaire

Please rate the degree to which you agree or disagree with each of the following, from 1 to 5 (where 1 - *strongly agree*, 5 - *strongly disagree*):

Cognitive load

(1) The learning content was difficult for me.

(2) I had to put in a lot of effort answering questions after the training.

(3) I found it difficult to answer questions after the training.

(4) I felt frustrated answering questions after the training.

(5) I did not have enough time to answer questions after the training.

Cognitive efforts

(6) I felt like I had a cognitive load while learning.

(7) I need to put a lot of effort to meet learning objectives and achieving learning goals.

(8) The teaching methodology was difficult to understand.